The Spine Functional Index (SFI): development and clinimetric validation of a new whole-spine functional outcome measure

ABSTRACT

BACKGROUND CONTEXT: Most spine patient-reported outcome measures are divided into neck and back subregions. This prevents their use in the assessment of the whole spine. By contrast, whole-spine patient-reported outcome measures assess the spine from cervical to lumbar as a single kinetic chain. However, existing whole-spine patient-reported outcomes have been critiqued for clinimetric limitations including concerns with practicality.

PURPOSE: To develop the Spine Functional Index (SFI) as a new whole-spine patient-reported outcome measure that addressed the limitations of existing whole-spine questionnaires; then to determine the SFI's clinimetric and practical characteristics concurrently with a recognised criterion, the Functional Rating Index (FRI).

STUDY DESIGN/SETTING: Observational cohort study within ten physical therapy outpatient clinics.

PATIENT SAMPLE: Spine-injured patients were recruited from a convenience sample referred by a medical practitioner to physical therapy. A pilot study (n=52, 57% female, age 47.6±17.5) followed by the main study (n=203, 48% female, age 41.0±17.8) that had an average symptom duration of less than five weeks.

OUTCOME MEASURES: SFI, FRI and Numerical Rating Scale (NRS).

METHODS: The SFI was developed through three stages: 1) item generation, 2) item reduction with an expert panel and patient focus group, then 3) pilot field testing to provide provisional clinimetric properties, sample size requirements and to determine suitability for a larger study. Participants completed the SFI, FRI and NRS every two weeks for six weeks, then every four weeks until discharge

or study completion at six months. Responses were assessed to provide individual psychometric and practical characteristics for both patient-reported outcomes, with the overall performance evaluated by the Measurement of Outcome Measures and Bot clinimetric assessment scales.

RESULTS: The SFI demonstrated high criterion validity with the FRI, (Pearson's r=0.87, 95%CI), equivalent internal consistency (α=0.91) and a single-factor structure. The SFI and FRI demonstrated suitable reliability ($ICC_{2,1}$=0.97:0.95), responsiveness (standardized response mean=1.81:1.68), minimal detectable change with 90%CI (6.4%:9.7%), Flesch-scale reading ease (64%:47%) and user errors (1.5%:5.3%). The clinimetric performance was higher for the SFI on the Measurement of Outcome Measures (96%:64%) and on the Bot scale (100%:75%).

CONCLUSIONS: The SFI demonstrated sound clinimetric properties with lower response errors, efficient completion and scoring, and improved responsiveness and overall clinimetric performance compared to the FRI. These results indicated that the SFI was suitable for functional outcome measurement of the whole-spine in both the research and clinical settings.

INTRODUCTION

Patients with pain or symptoms that arise from the spine may be evaluated with patient-reported outcome measures to determine their functional status [1-3]. These patient-reported outcome measures can be regional, designed to assess a region of the body, or the patient-reported outcome can be specific to a single joint, condition or disease. When assessing the functional status of patients with musculoskeletal conditions of the upper or lower limbs, a regional patient-reported outcome measure may be preferred as practicality is improved without compromising the essential psychometrics properties [4, 5]. However, when assessing the spine, patient-reported outcome measures remain distinctly divided into back [2] and neck [6]. Few whole-spine patient-reported outcome measures are recommended due to documented problems with either or both the psychometric and practical characteristics [2]. Another measurement option is a generic patient-reported outcome, such as the Short Form 36 Health Survey (SF-36) or the EuroQol. These generic patient-reported outcomes can be applied to all types of patients, regardless of their diagnosis or health problem [1]. However, these generic patient-reported outcomes have demonstrated reduced responsiveness over time because they do not contain sufficient items which are specific to the region, joint, condition or disease being assessed [7]. Consequently, these generic tools are less suited to measure regional musculoskeletal conditions [4, 5], including spine related conditions for both the back [4] and neck [8].

The adoption of the single kinetic chain concept for whole-spine patient-reported outcomes was first proposed by Williams et al [9]. Justifications supporting this concept included`: 1) patho-physiological grounds - as the aetiology for many mechanical non-specific spinal problems remains unknown; 2) co-existing regions - as presenting symptoms often occur in multiple, interconnected spinal areas; and 3) improved practicality - as one tool would provide measurement for all spinal areas [5, 10]. It has been recommended that a whole-spine patient-reported outcome be developed, particularly one that demonstrates acceptable clinimetric properties and performance, and subsequently compared to specific subregion spine patient-reported outcomes for the back and neck [9-11]. The

development and validation of a new whole-spine patient-reported outcome requires two phases: 1) initial development and evaluation of clinimetrics that includes concurrent validation with an existing whole-spine patient-reported outcome; then 2) subsequent concurrent validation with advocated criteria in separate subregions and condition-specific back and neck populations. This study's purpose was phase 1.

There are at least 43 back-specific patient-reported outcomes with 13 that can be used to evaluate responsiveness to change [2]. Among these the Oswestry Disability Index and Roland Morris Disability Questionnaire are the most commonly advocated [2, 12]. For the neck, at least 13 patient-reported outcomes have been developed [13] but there is limited agreement on which ones should be advocated [6, 8]. Five patient-reported outcomes purport validity for the whole-spine: the Functional Rating Index (FRI) [10], the Bournemouth Questionnaire [14], the Extended Aberdeen Spine Pain Scales [9], the Pain Disability Questionnaire [12] and the Core Outcome Measures Index [3]. However, further testing is required of these whole-spine tools because none have demonstrated an adequate factor structure through either Rasch analysis or factorial analysis [2], and the capacity to measure the whole-spine as a single kinetic chain [15]. Of these five patient-reported outcomes, the FRI is advocated most strongly due to its preferred administrative practicality and level of independent research on comparative clinimetric properties for both low back pain [16] and neck pain [8]. Consequently, the FRI is the optimal choice as a criterion measure ahead of the other four available whole-spine patient-reported outcomes when developing a new whole-spine patient-reported outcome and in preference to generic patient-reported outcomes such as the SF-36 or EuroQol.

The development of each of these five whole-spine tools has attempted to address the need for a single whole-spine tool. The initial three were questioned due to poor methodology in development, practicality, factor analysis and validation [2]. For example, the Pain Disability Questionnaire is not spine specific, nor does it account for acute situations as it is for 'chronic disabling musculoskeletal disorders' [12]. The eleven-item Core Outcome Measures Index has separate neck and back versions,

and is designed to measure patients after operative procedures within secondary and tertiary settings. Completion involves several scoring techniques with computerized input [3] and independent validation as a whole-spine measure is still required. Both the Aberdeen [9] and Pain Disability Questionnaire [12] have dual-factor structures that limit their validity as a single summated score and consequently are less than optimal measure [15]. The remaining three patient-reported outcomes have even less research in this aspect as they have not had their factor structure determined by the recommended maximum likelihood extraction method [17]. Consequently, a whole-spine patient-reported outcome is needed that has been appropriately developed [18], represents a single kinetic chain, has a single factor structure and appropriate clinimetric properties for both the back and neck.

A patient-reported outcome must be clinically practical, effective, efficient and validated with a recognized criterion standard [19]. The Spine Functional Index (SFI) (Figure 1) was developed to comply and satisfy these requirements. The aim of this study was to describe the development of the SFI, determine the psychometric, practical and factor structure characteristics in a general spinal population and compare the SFI to a whole-spine criterion measure, the FRI [10].

MATERIALS AND METHODS

A prospective observational study was completed in two phases (Figure 2):

1. SFI development in three-stages;
2. SFI validation in a symptomatic spine cohort.

Phase 1 - Development of the Spine Functional Index

The established three-stage development process used 1)item generation, 2)item reduction and 3)field testing [15, 18] (Figure 2).

Stage 1 Item Generation

Electronic data bases, PubMed, Cinahl, Embase and Pedro, from 1980-2010 were searched by the primary author (CPG) with key words ‘outcomes’, ‘self-report’, ‘function’, ‘disability’, ‘impairment’,

'spine', 'neck', 'back', 'thoracic', 'cervical' and 'lumbar'. An additional search included clinicians and researchers for unpublished questionnaires. This produced 129 patient-reported outcomes. A four-person peer-panel was formed, consisting of an occupational therapist, physical therapist with spine-specific post-graduate qualifications, general practitioner physician and occupational medicine physician with spine-specific consultancy work. The panel used consensus opinion that required a three vote minimum [20, 21] to review and shorten the list to 29 patient-reported outcome tools with 850 items that were directly cited in each of the patient-reported outcomes and relevant to the spine injuries. The list was reduced to 409 items by the panel through binning and winnowing methodology which removed duplicate and non-applicable items [22, 23].

Stage 2 Item Reduction

The 409 items were reduced in five separate stages (2a-e) by the panel. Stage 2a reduced the list to 159 items by pooling items with a common construct (e.g. 'sitting', 'sit in a chair', 'sit on a stool' etc. were collapsed to 'sitting'). Stage 2b classified [18] items using the World Health Organisation-International Classification of Functioning (WHO-ICF) [24] codes from the ICF Browser [25]: b=body functions, s=body structures, d=activities and participation, e=environmental factors [26]. Stage 2c reduced the 159 items to 89 by combining the ICF codes to common descriptive construct titles (e.g. 'stairs' and 'ladders' became 'code d4551-climbing'). Stage 2d reduced the list to 74 by grouping and deletion (e.g. 'dressing' and 'putting on pants' were retained but 'fastening clothing' was deleted). Stage 2e further combined items via consensus of importance and relevance to achieve the final 25 items, 15 general and ten spine-specific. The stems for each question were formulated: 'Due to my spine: I have difficulty/problems…'; or 'I stay/change/avoid/get others...'.

To ensure current best practice epidemiological standards were met, each question's final wording was achieved through peer panel consensus then given to two focus groups for feedback and relevance for face and content validity [18]: a spine symptoms patient focus group (n=10, three

cervical, three thoracic and four lumbar); the four person author group that included a physical therapist and an orthopaedic surgeon both with extensive experience in the spine, a biomechanist, and a physical therapist with extensive clinimetric research experience. The ten person patient focus group and the four person author panel supplemented the initial item reduction process performed by the 'expert panel'. The focus groups were provided with the final 25 items list and the list of the 49 items excluded in stages 2d. The mixed methods semi-structured interview process [27] was used to determine if the 25 items should be changed and if any of the 49 excluded items should be reinstated or included within the final item list. The "Isikawa" qualitative methodological process [28] was used to supplement the consensus agreement from both the patient and author focus groups and the expert panel. The format and three-item response option, 'Yes', 'No' and 'Half' [15], [29] were selected.

Stage 3 Field Testing

A pilot investigation enrolled 52 participants who provided a total of 85 responses (n^R). This ensured n=52 baseline responses and an additional 33 responses: 13 for reliability (n=13; n^R=26); and 20 for responsiveness, where two participants completed an additional third set of responses (n=18; n^R=38) (Figure 2). This allowed for a preliminary assessment of floor and ceiling effects, sampling method practicality and sample size calculations.

Sample Size

From the pilot study, minimum samples were determined for an 80% chance of detecting actual difference with 15% attrition (p<0.05) [30]. This compared favorably to previous FRI investigations [10, 31] for concurrent validity (n=106), reliability (n=56), responsiveness (n=84) and predictive ability through construct validity (n=168).

Phase 2 - Validation of the SFI in a cohort population

Design

A single stage, prospective observational study analyzed concurrent SFI and FRI responses. Each participant was classified by subregion (cervical, thoracic or lumbar) where the percentage noted ensured proportional reliability and responsiveness representation [15, 18].

Setting and Participants

Participants who complained of spinal pain or symptoms (n=203, responses=506) were consecutively recruited from ten Australian physical therapy clinics. Inclusion criteria were referral by a medical practitioner for a musculoskeletal spine condition or symptoms. Exclusion criteria were pregnancy, red flag signs, <18 years and English language difficulty. Symptoms and classifications of spinal diagnoses represent the entire spinal region, as described in Table 1.

Participants completed both the SFI and FRI patient-reported outcomes, however the number of FRI responses (n=173; responses=386) was reduced due to a misunderstandings with one participating clinic that returned only the SFI responses. Participants receiving ongoing treatment were re-measured every two weeks for six weeks, then every four weeks until discharge. Status was classified as: acute at 0-6 weeks; subacute at >6-12 weeks; and chronic at >12 weeks. Pooled responses assessed criterion validity, distribution and missing responses. Participants also completed an 11-point global numeric rating scale of perceived present overall status [32, 33], whereby subjects rate their status on a scale from 0-10 (0=worst possible, 10=normal). The global numeric rating scale was used as an external criterion measure of clinical change, by calculating the difference in global perceived present status over time.

Questionnaires

The FRI [10] is a single page patient-reported outcome that contains 10 items, each rated on a five-point Likert scale incorporating visual and descriptive response options. Five items on the FRI are common to the Oswestry Disability Index and the Neck Disability Index with three additional Oswestry Disability Index items, one Neck Disability Index item and a new ‘pain’ item [2]. The raw score of the

FRI is multiplied by 2.5 to generate a 0-100% score on the FRI (100%=no disability). One missing response is permitted.

The SFI is a single page 25-item patient-reported outcome, with a three-point Likert scale response option for each item. The scores from the 25 items are tallied for the sum, the sum is multiplied by four and then subtracted from 100 to generate a 0-100% score (100%= no disability). Two missing responses are permitted.

An 11-points global numerical rating scale (0=worst possible, 10=normal or fully recovered) was used to reflect individual perceived global functional status and act as an external criterion.

Data Analysis - Psychometric Characteristics

Distribution and normality were assessed from baseline histogram inspection and one-sample Kolmogorov-Smirnov tests (significance >0.05) [30]. *Internal consistency* used baseline Cronbach's Alpha (α=0-1.00) calculations with an optimal value recommended as 0.90-0.95 [18, 30]. *Test-retest reliability* was assessed through the Intraclass Correlation Coefficients (ICC) Type 2,1, and expressed with 95% CI using scores on the patient-reported outcome from acute/subacute participants at baseline and again on day three during a non-treatment period. Participants rating on the global numerical rating scale of perceived overall status at baseline and on day three provided the reference criterion to determine change. Only those participants who had a change of 0, +/-1 were entered into analysis for test-retest reliability (n=70) [15].

Responsiveness was assessed using the effect size and standardized response mean statistics [18]. Participants were classified by subregion with repeated measures analyzed (n=191 for the SFI; n=144 for the FRI) for: acute at two weeks, subacute at four weeks and chronic at six weeks. This accounted for variations in healing and therapists interventions [15]. There were participants that received no follow-up or early discharge (SFI, n=12; FRI, n=7). The global numeric rating scale score of a change $\geq$2.0 was the cut-off used to define patient-rated clinical change. *Error score* was determined with the

minimal detectable change with 90% confidence bounds (MDC_{90}) using the standard error of the measurement formula and the ICC coefficients. *Minimal clinically important difference (MCID)* was calculated using an anchor based method, with the anchor of patient-rated change determined from the global numeric rating of change. Patients were classified as improved or deteriorated if they had a minimum change of $\geq$2.0 points on the global numeric rating scale between baseline and follow-up [18, 33, 34]. Consequently, the MDC appears as a statistically and clinically appropriate MCID [35].

Validity was assessed for *face* and *content* through focus groups, panel feedback and readability scores [36]; and *criterion* through Pearson's r coefficient (n=386). *Construct validity* used discriminant validity with the external criterion global numeric rating scale of perceived self-rated change of health status $\geq$2.0 points [34]. Additionally, an *a-priori* paired t-test statistical difference was required between baseline and repeated test groups mean scores to categorize subjects as improved or deteriorated when calculating the MCID. *Factor analysis* used baseline SFI and FRI data with loading suppression at 0.30 and varimax rotation for maximum likelihood extraction [30] which required assumptions of normality. Factor extraction had three *a-priori* requirements: scree-plot 'point of inflection'; eigenvalue >1.0; and variance $\geq$10% [30].

Data Analysis - Practical Characteristics, Readability and Summary Performance

Practicality considered nine areas [15, 36] with being five self-evident: 1) self-administered; applicable across a variety of 2) conditions; 3) severity levels; 4) relevance to defined populations; and 5) single-page length. The remaining four areas were determined through focus groups for 6) interviews for ease of understanding and completion; 7) questionnaire completion time; 8) therapist scoring-time from three separate scores averaged from each clinic; and 9) missing responses as percentages of total responses (SFI, n=506; FRI, n=386). *Readability* used the Flesch–Kincaid grade scales (range 0-12, optimum<7) and reading-ease (optimum>60%) calculated from word-processing software. Summary performance used the 'Measurement of Outcome Measures' scale that evaluated 25 essential properties

[5]; and the 'Bot' scale that evaluated twelve items [36]. The 'Bot' cut-off classifications were adjusted [15, 29] for 'time to administer' at three minutes and 'readability and comprehension' determined by the Flesch-Kincaid scale cut-offs [15]. Significance was set at $p<0.05$.

RESULTS

Participant demographics are reported in Table 1.

Psychometric Properties

Characteristics of internal consistency, reliability, responsiveness and error score are summarized in Table 2, and construct validity in Table 3.

Distribution and normality were demonstrated through the Kolmogorov-Smirnov test (SFI=1.163, significance=0.87; FRI=1.18, significance=0.87) with identical SFI and FRI baseline score ranges (0% -98%). The SFI histogram shape was preferred particularly in the upper 90-100% interval that contained 15 (7.5%) responses compared to the FRI with a single response (2%). The 'Half Mark' option was used by 57% of participants at baseline and in 43% of all responses. The baseline scores by subregion were comparable between the SFI and FRI apart from the multi-area group (Table 4).

Criterion validity was high (Pearson's r=0.85) between the SFI and FRI scores. *Construct validity* through *discriminant validity* was demonstrated for the *a-priori* criterion (Table 3). The subregion mean scores were different for both patient-reported outcomes and between both patient-reported outcomes, though the cervical, thoracic and multi-area groups were of a similar value. However, none were statistically significant apart from the multi-area group ($p<0.001$).

Factor analysis was suitable as the correlation matrix Kaiser-Meyer-Oklin value was 0.912 and Barlett Test of Sphericity significant ($p<0.001$). A unidimensional structure was indicated for both patient-reported outcomes as the three *a-priori* criteria were met with second point scree-plot inflection

and one eigenvalue >1.0 where variance was >10% (SFI=33.4%, FRI=55.6%). The SFI had six more factors with eigenvalues >1.0, but with variance <10% that accounted for 30.5%. Both patient-reported outcomes had four factors with eigenvalues between 0.5 and 1.0 with the remaining factors all below a 0.5 eigenvalue.

Practical Characteristics

Completion time was SFI=122±37 seconds, and FRI=84±23 seconds; scoring time was SFI=16±4 seconds, FRI=27±13 seconds. The FRI required a computational aid, and with one missing response the scoring time increased to 53±19 seconds. Combined completion and scoring was SFI=138±41 seconds, FRI=137±39 seconds. *Missing responses* were <1.5% for the SFI, and 5.3% for the FRI. *Readability* for the SFI was grade=7, reading ease=64%; and for the FRI grade =7, reading ease =47.2%. *Summary performance* on the Measurement of Outcome Measures was SFI=96%, FRI=64%; and on the 'Bot' for the SFI=12/12 or 100%, FRI=9/12 or 75%.

DISCUSSION

The SFI was developed using a structured methodology. It demonstrated acceptable psychometric properties, a single factor structure, and strong practical characteristics in patients with spinal pain and symptoms of the cervical, thoracic and lumbar spine. Compared to the FRI, by visual comparison of the results, the SFI had equal or preferable psychometric properties of reliability, validity, responsiveness and error. The summary performance scores of practical characteristics on the Measurement of Outcomes Measure and 'Bot' scales showed high scores for the SFI. The SFI was demonstrated to be capable of assessing functional status at a single point in time and change over time to determine the effectiveness of treatment interventions. The practical characteristics of short scoring times, low missed responses and reading ease will reduce both the patient and administrative burden.

The SFI has a three-point response format, which was used by participants 57% of the time. This response format provided a simple scoring format within a stable equally-spaced scale [37]. This also enabled sound individual interpretation for the psychological perspective of an item's 'presence', 'absence' or an 'intermediate position' [38] as opposed to a dichotomous response option.

Normalised SFI distribution and subregion scores in this cohort of patients presenting to physical therapy demonstrated no floor or ceiling tendency. The FRI had more missing responses at the higher levels of functional loss, indicating reduced measurement capacity. This measurement capacity of the SFI may improve the ability to discriminate change throughout the scale range. Internal consistency, test-retest reliability and responsiveness values for the SFI were acceptable and comparable to the FRI. The SFI demonstrated lower error values (SEM and MDC_{90}) which may allow for improved sensitivity for detecting change over time in the assessment of intervention effectiveness that may otherwise not show a valid effect [39]. Moreover, this may subsequently reduce the 'number needed to treat' [40].

Responsiveness of the SFI in a cohort of patients undergoing physical therapy treatment was acceptable and comparable to the FRI, despite the higher diversity in baseline impairment [41, 42]. As an observational study in a cohort of patients undergoing physical therapy care, other influences on responsiveness may have been present. These include variation in interventions provided, follow-up duration (as responsiveness is less over a shorter follow-up period) and baseline severity (as acute and chronic patients change at different rates) [34]. These variables were attempted to be minimize by using the concurrent testing methodology. Factor analysis demonstrated a single-factor structure and consistent variance levels for both the SFI and FRI. This study is the first to report the FRI factor structure.

Limitations and Strengths of the Study

One limitation of this study was the recruitment of patients presenting for care at physical therapy outpatient clinics only. Consequently, results cannot be generalized to inpatient or community settings.

Patients referred to physical therapy most likely represent the mid-range of spine conditions. The study's strengths were the prospective, multi-center investigation that included patients from each spinal region with varied degrees of severity and duration that represented both the general and work injured populations with a large variation in diagnoses (Table 1). Furthermore, 191 subjects were available for the responsiveness sample, measuring these subjects on repeated occasions over time. This facilitated their measurement throughout the severity spectrum, as indicated by the suitable levels of distribution within the histogram, including the least affected level at the point of discharge.

Implications for Further Research

The high SFI and FRI criterion validity implied generalizability to populations where the FRI has been validated or compared to other spine related patient-reported outcomes. `This includes the Oswestry Disability Index, Roland Morris Disability Questionnaire and Neck Disability Index. However, independent investigations are required where spine subregion patient-reported outcomes are concurrently compared through repeated measures on diagnoses such as whiplash, acute and chronic low back and neck pain. The SFI had several factors that accounted for substantial variance. This suggests that shortening to perhaps ten-items may be possible. This may improve practicality and reduce both respondent and clinician burden. A confirmatory factor analysis should be considered.

CONCLUSIONS

The SFI is a practical patient-reported outcome for measurement of spine related patient status and change over time. Compared to the FRI, an advocated whole-spine patient-reported outcome, the SFI had comparable and sometimes improved psychometric and practical characteristics and overall performance. The findings of this study indicated the SFI is a viable patient-reported outcome for measuring whole-spine functional status in both the clinical and research settings.

List of References

1. Garratt, A., *Patient reported outcome measures in trials, Editorial.* BMJ, 2009. 338(a): p. 2597
2. Cleland, J.A., R. Gillani, E. J. Bienen, and A. Sadosky, *Assessing Dimensionality and Responsiveness of Outcomes Measures for Patients with Low Back Pain.* Pain Pract, 2011. 11(1): p. 57-69..
3. Mannion, A.F., F. Porchet, F. Lattig, et al., *The quality of spine surgery from the patient's perspective: part 2. Minimal clinically important difference for improvement and deterioration as measured with the Core Outcome Measures Index.* Eur Spine J, 2009. 18(Suppl 3): p. 374-9..
4. Garratt, A.M., M.J. Klaber, and A.J. Farrin, *Responsiveness of generic and specific measures of health outcome in low back pain.* Spine, 2001. 26(1): p. 71-7.
5. Gabel, C.P., L. Michener, B. Burkett, and A. Neller, *The Upper Limb Functional Index (ULFI): Development and Determination of Reliability, Validity and Responsiveness.* J Hand Ther, 2006. 19(3): p. 328-349.
6. van der Velde, G., D. Beaton, S. Hogg-Johnston, et al., *Rasch analysis provides new insights into the measurement properties of the neck disability index.* Arthritis Rheum, 2009. 61(4): p. 544-551.
7. Suarez-Almazor, M.E., C. Kendall, J. A. Johnson, et al., *Use of health status measures in patients with low back pain in clinical settings. Comparison of specific, generic and preference-based instruments.* Rheumatology (Oxford), 2000. 39(7): p. 783-90.
8. Rebbeck , T., D. Sindhusake, I. D. Cameron, et al., *A prospective cohort study of health outcomes following whiplash associated disorders in an Australian population.* Inj Prev, 2006. 12(2): p. 93-8.
9. Williams, N., C. Wilkinson, and I.T. Russell, *Extending the Aberdeen Back Pain Scale to include the whole spine: a set of outcome measures for the neck, upper and lower back.* Pain, 2001. 94(3): p. 261-274.
10. Feise, R.J. and J.M. Menke, *Functional Rating Index. A new valid and reliable instrument to measure the magnitude of clinical change in spinal conditions.* Spine, 2001. 26(1): p. 78-86.
11. Gabel, C.P., B. Burkett, and M. Yelland, *Balancing Fidelity and Practicality in Short Version Musculoskeletal Outcome Measures* Phys Ther Rev, 2009. 14(4): p. 221-225.
12. Anagnostis, C., R.J. Gatchel, and T.G. Mayer, *The pain disability questionnaire: a new psychometrically sound measure for chronic musculoskeletal disorders.* Spine, 2004. 29(20): p. 2290-302; discussion 2303.
13. Resick, D., *Subjective outcome assessments for cervical spine pathology: A narrative review.* J Chiro Med, 2005. 3(4): p. 113-134.
14. Bolton, J.E. and B.K. Humphrey's, *The Bournemouth Questionnaire: a short-form comprehensive outcome measure. II. Psychometric properties in neck pain patients.* J Manipulative Physiol Ther, 2002. 25: p. 141-148.
15. Gabel, C.P., L. Michener, M. Melloh, and B. Burkett, *Modification of the Upper Limb Functional Index to a Three-point Response Improves Clinimetric Properties.* J Hand Ther, 2010. 23(1): p. 41-52
16. Chansirinukor, W., C. G. Maher, J. Latimer, J. Hush, *Comparison of the functional rating index and the 18-item Roland-Morris Disability Questionnaire: responsiveness and reliability.* Spine, 2005. 30(1): p. 141-5.
17. Fabrigar, L.R., D. T. Wegener, R. C. MacCallum, E. J. Strahan, *Evaluating the use of exploratory factor analysis in psychological research.* Psychol Methods, 1999. 4(2): p. 272-299.
18. Streiner, D.L. and G.R. Norman, *Health Measurement Scales: A Practical Guide to Their Development and Use*. 4 ed. 2008, Oxford: Oxford University Press.
19. Liang, M.H. and A.M. Jette, *Measuring functional ability in chronic arthritis: a critical review.* Arthritis Rheum, 1981. 24: p. 80-86.
20. Kirshner, B. and G.H. Guyatt, *A methodological framework for assessing health indices.* J Chronic Dis, 1985. 38(1): p. 27-36.
21. Kopec, J.A., E. C. Sayre, A. M. Davis, et al., *Assessment of health-related quality of life in arthritis: conceptualization and development of five item banks using item response theory.* Health Qual Life Outcomes, 2006. Jun 2(4): p. 33.

22. Kopec, J.A., *Measuring functional outcomes in persons with back pain: a review of back-specific questionnaires.* Spine, 2000. 25(24): p. 3110-4.
23. Patient -Reported Outcome Measurement Information Systems (PROMIS) (2010) *Version 1.0 Item Banks* Volume, http://www.nihpromis.org/science/ItemClassification, Accessed Aug 20 2011.
24. World Health Organisation (WHO). *International Classification of Functioning, Disability and Health (ICF)*. [web site] 2001 [cited 2009 July 31]; Available from: www.who.int/icidh.
25. World Health Organisation, W. *The ICF Browser*. [Internet] 2010 [cited 2010 January 12 2010]; Web Browser: http://apps.who.int/classifications/icfbrowser/].
26. Escorpizo, R., G. Stucki, A. Cieza, et al., *Creating an Interface Between the International Classification of Functioning, Disability and Health and Physical Therapist Practice.* Phys Ther, 2010. 90(7): p. 1053-63.
27. Johnson, R.B., A.J. Onwuegbuzie, and L.A. Turner, *Toward a Definition of Mixed Methods Research.* J Mix Methods Res, 2007. 1(2): p. 112.
28. Ishikawa, K. and J.H. Loftus, *Introduction to quality control*. 1990, Tokyo: 3A Corporation.
29. Gabel, C.P., M. Melloh, and B. Burkett, *The Lower Limb Functional Index: development and validation of the clinimetric properties and practical characteristics.* Phys Ther, 2012. 92(1): p. 98-110.
30. Field, A., *Discovering Statistics using SPSS*. 2nd ed. 2005, London: SAGE Publications Ltd.
31. Childs, J.D. and S.R. Piva, *Psychometric properties of the functional rating index in patients with low back pain.* Eur Spine J. 2005, 2005. 14(10): p. 1008-12.
32. Bowling, A., *Just one question: If one question works, why ask several?* J Epidemiol Community Health, 2005. 59(5): p. 342-5.
33. Ostelo, R.W., R. Deyo, P. Stratford, et al., *Interpreting change scores for pain and functional status in low back pain: towards international consensus regarding minimal important change.* Spine, 2008. 33(1): p. 90-4.
34. Childs, J.D., S.R. Piva, and J.M. Fritz, *Responsiveness of the numeric pain rating scale in patients with low back pain.* Spine, 2005. 30(11): p. 1331-4.
35. Copay, A., S. D. Glassman, B. R. Subach, et al., *Minimum clinically important difference in lumbar spine surgery patients: a choice of methods using the Oswestry Disability Index, Medical Outcomes Study questionnaire Short Form 36, and pain scales.* Spine J., 2008. 8(6): p. 968-74.
36. Bot, S.D., C. B. Terwee, D. A. van der Windt, et al., *Clinimetric evaluation of shoulder disability questionnaires: a systematic review of the literature.* Ann Rheum Dis, 2004. 63(4): p. 335-41.
37. Krosnick, J.A., *The handbook of questionnaire design* 1991, New York: Oxford University Press.
38. Albarracin, D., B.T. Johnson, and M.P. Zanna, *The Handbook of Attitudes*. 2005, Hillsdale, NJ: Erlbaum.
39. Stratford, P.W. and D.L. Riddle, *Assessing sensitivity to change: choosing the appropriate change coefficient.* Health Qual Life Outcomes, 2005. 3: p. 23.
40. Moore, R.A., S. S. Smugar, H. Wang, et al., *Numbers-needed-to-treat analyses – Do timing, dropouts, and outcome matter? Pooled analysis of two randomized, placebo-controlled chronic low back pain trials.* Pain, 2010. 151(3): p. 592-597.
41. Cohen, J., *Statistical power analysis for the behavioral sciences*. 1988, Hillsdale, NJ: Erlbaum.
42. Liang, M.H., A.H. Fossel, and M.G. Larson, *Comparison of five health status instruments for orthopaedic evaluation.* Med Care, 1990. 28: p. 632-642.

FIGURE LEGENDS

Figure 1: Spine Functional Index

Figure 2: Flow chart of SFI development and validation